特集

ナレッジマネジメントを支える文書処理技術

Document Processing Technology for Knowledge Management

あらまし

企業においては，市場やニーズの変化に即応した活動が必要であり，個々人がばらばらに所有している様々な知識を企業の構成員間で共有し活用することが課題となっている。このように人が知識をうまく活用すること（ナレッジマネジメント）は，究極は人の知的能力の問題であるが，人が知的能力を最大限に発揮するために，情報システムにより支援できる部分も多い。

人が知識を活用するためには，知的活動の個々のプロセスの中で，多くの情報の中から必要な情報を見つけ出して，付加価値を与えたり，人とのコミュニケーションにおいて，新たな知識を創造することが重要である。この中で，多くの情報の中から必要な情報を見つけ出すことが中心課題である。

本稿では，ナレッジマネジメントを支援する技術の中で，基盤となる文書処理に焦点を絞り，情報検索，要約，情報抽出，分析などの技術について，研究例を交えて紹介する。

Abstract

Society is becoming more complex, and various new types of communication and processing have become essential. As a result, it has become more important for us to share and use our knowledge effectively. It is especially important for corporations to respond quickly to changes in the market and users 'needs. Knowledge possessed by individuals should be shared between members of corporations and used. Such knowledge sharing and use is called knowledge management, and its success depends on the intellectual capabilities of individuals. These capabilities can be more fully exploited by using information processing systems that support the various aspects of knowledge management.

Obtaining and using knowledge often involves sifting through a large amount of unneeded information and then, after obtaining the needed information, adding new value to it through many intellectual processes. It is also important to create new knowledge during personal communication. The most basic of these activities is finding information from among a large amount of unneeded information.

In this paper, we mainly explain document processing, which is a basic technology for supporting knowledge management. We introduce information retrieval, summarization, extraction, and analysis technologies and various examples of research into these technologies.

山本栄一郎（やまもと　えいいちろう）

1973年東京工業大学工学部電子工学科卒。同年(株)富士通研究所入社。以来文字認識，画像処理，人工知能，ヒューマンインタフェース，ドキュメント処理の研究開発に従事。コンピュータシステム研究所ドキュメント処理研究部

ま え が き

社会が複雑になり，多様な対応が必要になってくるにつれ，個々の人々が経験的に獲得してきた知識をお互いに共有し，活用することが重要になってきた。そのため，知識をうまく活用するための枠組みとして，ナレッジマネジメント(以下，KM)が，多くの分野において注目されている。

とくに企業においては，変化の速い市場やニーズに即応した活動なくしては，持続的な発展を続けることが難しくなってきている。そのため，企業が所有している様々な知識を企業の構成員が共有し活用することにより，知的生産性を向上させることが至上の課題となっている。

KMは，人の知的能力の問題であるが，人が知的能力を最大限に発揮できるようにするために，情報システムが支援できる部分は大きい。

本稿ではKMを支援する技術として，その基盤となる文書処理技術に焦点をあて，情報検索，要約，情報抽出，分析などの研究開発技術について紹介する。

KM を支援する技術

KMを支援するシステムには，大きく分類して，つぎの３種類のことが要請される。

(1)　大量の情報からの知識発見支援

多くの場合，求める知識は，新聞記事データベースや企業内の情報格納庫(Webなど)のような大量の情報の中に埋もれている。したがって，大量の情報の中から，利用者の欲しい知識を容易に見つけ出すことを支援する必要がある。

(2)　業務プロセスに沿った支援

ある業務を考えると，その業務の特定のステップで必要とされる知識は，一般的な知識ではなく，かなり限定されたものである。したがって，その業務プロセスを分析し，各ステップで必要とされる知識の種類をそのステップに関連付けておけば，すべての情報を一括して蓄積する場合に比べて，知識の利用や生成が容易になる。

(3)　コミュニケーションの支援

大量に存在する情報から，有効な情報を見出し，いくつかの情報を組み合わせるなどして，情報を知識として活用するのは，最終的には人である。人と人とのコミュニケーションは，新しい知識を生み出す契機となる。したがって，ネットワーク上にコミュニケーションの場を提供するなど，コミュニケーションを支援することが重要である。

さて，企業内やグループ内で知識を共有し，それを有効に活用するためには，上記３種類の要請の中で，まず，(1)項がKMを支援するシステムの基盤となる。

文書処理技術

● 情報検索

KMでは，企業内やグループ内に蓄積された大量の文書を利用するため，まず第一に，大量の文書の中から，利用者の欲する情報が含まれた文書を高速に検索できることが重要である。著者らは，すでに大容量情報全文検索エンジンTerass[(1)]を開発し，InfoNavigatorなどの商用検索サービスの中でも用いている。

ところで，検索語を指定することにより，利用者の欲しい文書を大量の文書の中から検索する場合，検索語が適切でないと，大量の文書が検索され，その中から利用者が本当に欲しい文書を取り出すのは，非常に骨が折れる。また，一般の人にとって，適切な検索語を思いつくのも容易なことではない。

これを解決するため，インターネットなどにおいては，利用者が，用意されたディレクトリをクリックしていくだけで，目的の文書にたどり着ける文書ディレクトリを備えているものもある。しかし，一般に，文書ディレクトリは人手で構築されているため，維持コストが非常に大きい。そこで，ディレクトリの半自動構築システ

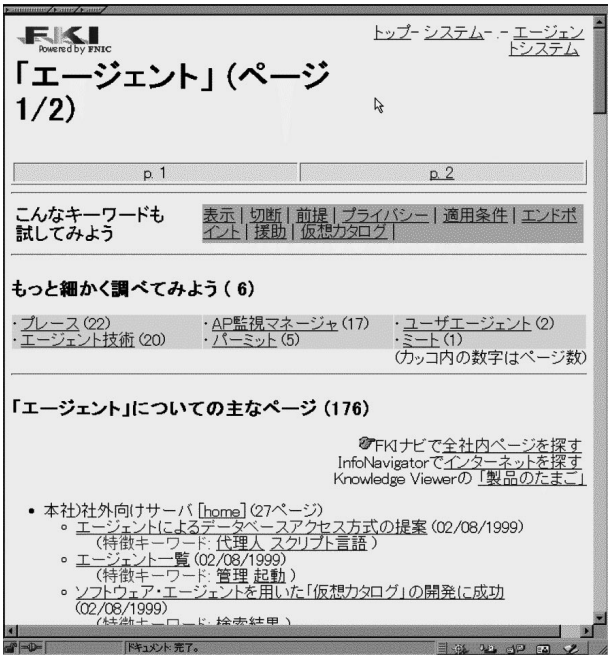

図-1　ディレクトリの画面表示例
Fig.1-Screen image example of document directory.

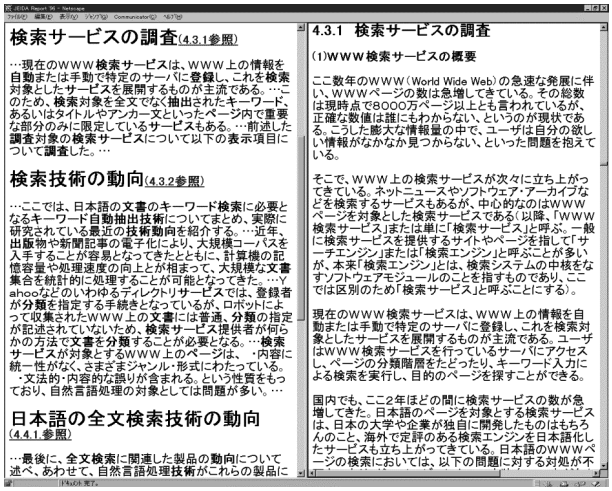

図-2　要約の例（左側：要約，右側：原文）
Fig.2-Example of summarization.

ムWINDを開発した。[2]

このシステムでは，まず，相関ルールの手法を用いて，キーワード集合間の階層・連想関係を抽出する。つぎに，キーワード集合に対して，階層・連想関係に基づくリンクを張ることで，文書ディレクトリを自動生成している。

社内イントラネットの約27万URLに適用し，約２万ノードのディレクトリを生成した。生成されたディレクトリの画面例を図-1に示す。この画面では，検索範囲を広げる上位キーワード，検索範囲を狭める下位キーワードやブラウジングを支援するための連想キーワードが表示されている。

● 要約

情報検索では，利用者の欲しい情報を含んだ文書の候補を取り出すことはできるが，その文書の中に，利用者が欲しい情報が含まれているかどうかは分からない。したがって，長大な文書の中に，利用者の欲しい情報が含まれているかどうかを容易に見分けられるように，長大な文書を要約する技術の研究開発を進めている。[3]

本技術では，

(1)　文書中から，すべての語彙を抽出
(2)　同一語彙の繰り返しを基に定義した「結束度」を算出することにより主要な話題を抽出
(3)　各話題のまとまりから重要語を抽出し，文章の重要箇所を推定
(4)　文章の重要箇所から主要な文を抜粋

という手順で，要約を作成している。図-2は，全体で81ページの文書（右側）を本手法により，１ページに要約（左側）した例である。

銘柄名	製品種	製品名	価格	発売日	記事見出し
高千穂酒造	高濃度の高級焼酎（しょうちゅう）		2,500	11/12/96	長期貯蔵焼酎2製品、高千穂酒造が
ニッカウヰスキー	韓国産しょうちゅう			04/05/94	ニッカ、韓国産しょうちゅう発売
岩崎産業（鹿児島市、岩崎福三社長）グループ	芋焼酎（しょうちゅう）	同窓会シリーズ			岩崎産業グループ、高校名鑑
薩摩酒造	麦しょうちゅう	琥珀（こはく）の夢	1,600	03/09/94	薩摩酒造、ぎょう発売、樽で
島津興業	錦灘酒造（鹿児島県溝辺町、山元正博社長）と提携して米焼酎（しょうちゅう）	島津雨			島津興業が販売、上級
薩摩酒造	いもしょうちゅう	さつま白波・千支（えと）ラベル	1,200		薩摩酒造、「戌」ラベ
（鹿児島）田苑酒造	「いにしえの香り漂う焼酎（しょうちゅう）」をキャッチフレーズに、イモ焼酎	煮たて			高温発酵のイモ焼酎
宝酒造	本格純米しょうちゅう	よかいち	1,250	03/05/93	宝酒造、しょうちゅう限。
日露酒造	高級焼酎	季の詩（ときのうた）	2,000	03/28/92	日露酒造、トウモロコシ
舞酒造	びん入りの純米焼酎（しょうちゅう）	舞心		08/01/91	花の舞酒造、びん入り純米

図-3　情報抽出の例
Fig.3-Example of information extraction.

● 情報抽出

要約よりさらに簡潔な形で必要な情報を文書中から取り出すのが，情報抽出である。

利用者の欲しい情報が，ある一つの文書中に含まれているのではなく，複数の文書を調べて初めて得られる場合などに，情報抽出は有効である。

情報抽出の例を図-3に示す。この例は，新聞記事の中から，焼酎の製品に関する情報を抽出したものであり，どんな銘柄の焼酎が，どこから，いくらで発売されているかなどの情報を一覧することができる。

技術としては，あらかじめ特定の事象（新製品発売など）を表現する文章のパターンを抽出しておき，これを新たな文書に適用することにより，特定の事象を自動的に抽出している。[4]

この技術は，商用のデータベースサービス（ジー・サーチ）で実際に用いられている。

● 分析

自由回答形式のアンケート調査のような大量の文書を有効に利用するためには，そこに書かれている内容を自動的に分析し，利用者に分かりやすい形で提示する必要がある。

HIPS[5]は，このような目的のために研究開発を進めているもので，情報（単語，文書）間の関連性を抽出するKeyword Associatorと，関連情報の視覚的操作環境であるD-ABDUCTORから構成される。

Keyword Associatorは，単語の出現頻度を基に，単語や文書間の関連度を計算する。D-ABDUCTORは，様々なグラフの自動描画機能を備えており，Keyword Associatorが計算した単語間の関連度に従って，単語や文書の分布をビジュアルに表現する。

図-4は，パソコンに関するアンケート結果を分析・表示したものである。上下左右の４点に，それぞれ「デス

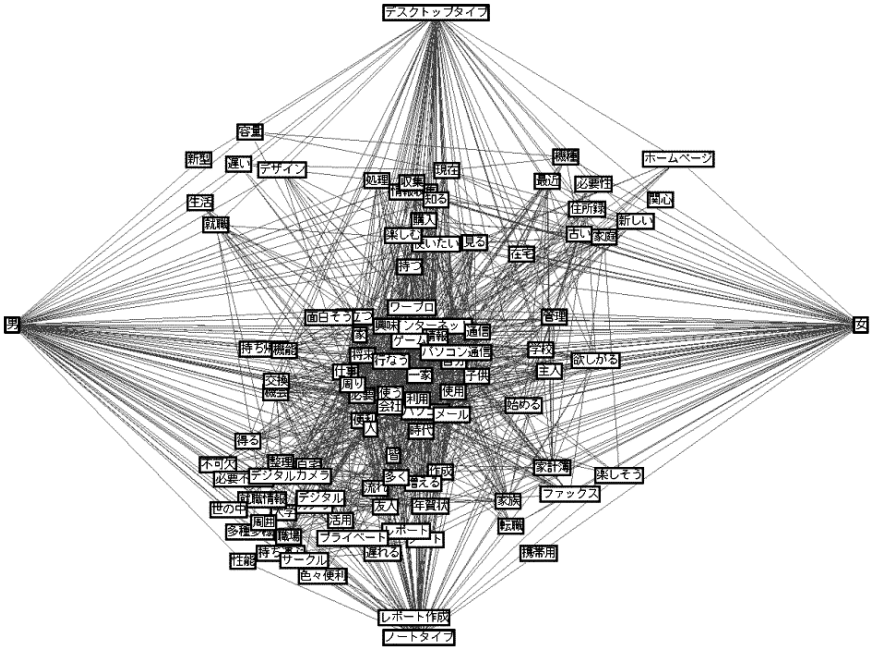

図-4　HIPSによる分析例
Fig.4-Example of analysis by HIPS.

クトップタイプ」「ノートタイプ」「男」「女」という単語を配置したとき，ほかの単語がどのような位置(関係)にあるかを示している。

む　す　び

本稿では，KMを支援するシステムの基盤となる文書処理技術について，研究例を交え紹介した。

情報に付加価値を見出し，知識として活用するのは，最終的には人である。しかし，情報システムによってKMを支援できる部分は，まだまだ残されている。人が「創造性」や「高付加価値化」などの「知的生産性」を向上させることを支援するため，今後も研究開発に取り組んでいく。

参考文献

(1)　松井ほか：大容量情報全文検索エンジン．*FUJITSU*，48，3，pp.240-243(1997)．

(2)　津田：テキストマイニングを用いた文書のディレクトリ整理．ソフトウェア科学会第15回大会，1998.

(3)　仲尾：文書の話題構成に基づく重要語の抽出．情処研報，FI50-1，1998.

(4)　西野ほか：トップダウンなパターン解析に基づく情報抽出．情処研報，NL124-13，1998.

(5)　渡部ほか：ハイブリッド発想支援システム「HIPS」．計測自動制御学会第17回システム工学部会研究会，pp.77-84(1995)．